Amendments to the Specification

Beginning at page 6 and continuing onto page 7, please amend Paragraph **[0022]** as shown below:

[0022] Figure 6 is a more detailed flow chart of one embodiment of the duplicate detection process. Step 601 initiates identification of potential duplicates. The particular embodiment depicted assumes that certain data is available. It assumes that nearest neighbors have already been identified, similarity scores calculated, and nearest neighbors ordered according to similarity scores. Step 602 is the beginning of a loop which iterates for each document in the database. Step 603 is the beginning of a loop which iterates for each nearest neighbor of a particular document. Step 604 is a test for whether the similarity score for a current nearest neighbor is equal to or within a predetermined threshold t of a next nearest neighbor. If this test fails, the process loops back to step 603. If the test 604 passes, step 605 determines whether the matching nearest neighbors are already found in among the identified pairs of potential duplicates. If they are, the process is shortcut and control returns to the top of loop 603. If not, the process continues to step 606. If the test 606 fails, the process loops back to step 603. Optionally, the pair are tested for whether they are top scorers for each. Two or more identical documents will have a much higher similarity ~~scores~~ **score** to each other than they will to any non-identical document. Accordingly, two or more identical documents, are likely to be top scorers for each other. The process proceeds to step 607. The test summarized by the **step** 607 is explained in greater detail in the sequence of steps 641-48. Overall, the issue is whether the pair of documents with equal or similar similarity scores actually match. If the feature indices vectors or other characteristics of the document pair do not match, the process loops to step 603. If they do match, the pair of documents are added to the list of found pairs in step 608. Adding the pair to the list of duplicate sets involves different steps, depending on whether both, one or neither of the documents is already in a duplicate set. Test 609 determines whether both of the pair of documents ~~is notT~~ **is not** already in a set of duplicates. If this test fails,

control passes to step 610. If it succeeds, control passes to step 621. Test 610 determines whether at least one of the pair are already in a set of duplicates. If this test fails, control passes to step 611. If it succeeds, control passes to step 631. In step 611, a new duplicate document set is created and the pair of documents are used to create a new set. Step 612 is the end of the loop that began with step 603. If all of the nearest neighbors of a particular document have been processed, flow proceeds to step 613; otherwise, it loops back to step 603. Step 613 is the end of the loop for processing a particular document or data set item. If there are more documents or data set items to process, flow loops to step 602. Otherwise, iteration through this flowchart is complete. After iteration is complete, or, alternatively, throughout the process, the user may be presented with an interface for responding to the detected potential duplicates.